

# Probability and Statistics for JMC

Lectured by Alex Geringer-Sameth

Typed by Aris Zhu Yi Qing

May 17, 2021

# Contents

<b>1</b>	<b>Review of Elementary Set Theory</b>	<b>3</b>
<b>2</b>	<b>Visual and Numerical Summaries</b>	<b>4</b>
2.1	Visualization . . . . .	4
2.2	Summary Statistics . . . . .	6
2.2.1	Measures of Location . . . . .	6
2.2.2	Measures of Dispersion . . . . .	7
2.2.3	Covariance, Correlation, and Skewness . . . . .	7
2.2.4	Box-and-whisker plot . . . . .	8
<b>3</b>	<b>Probability</b>	<b>9</b>
3.1	Formal Definition of Probability . . . . .	9
3.1.1	$\sigma$ -algebra . . . . .	9
3.1.2	Probability Measure . . . . .	10
3.2	Conditional Probability . . . . .	11
3.3	Independence . . . . .	12
3.4	Bayes' Theorem . . . . .	12
<b>4</b>	<b>Discrete Random Variables</b>	<b>14</b>
4.1	Random Variables . . . . .	14
4.2	Discrete Random Variables . . . . .	15
4.3	Functions of a Discrete Random Variable . . . . .	17
4.4	Mean and Variance . . . . .	17
4.5	Some important Discrete Random Variables . . . . .	19
<b>5</b>	<b>Continuous Random Variable</b>	<b>21</b>
5.1	Definitions . . . . .	21
5.2	Transformations . . . . .	22
5.3	Mean, Variance and Quantiles . . . . .	24
5.4	Some Important Continuous Random Variables . . . . .	25
<b>6</b>	<b>Jointly Distributed Random Variables</b>	<b>28</b>
6.1	Definitions . . . . .	28
6.2	Independence, Conditional Probability, Expectation . . . . .	30
6.3	Multivariate Transformations . . . . .	32

6.4	Gamma and Beta Distributions . . . . .	33
<b>7</b>	<b>Convergence Concepts and Theorems</b>	<b>35</b>
7.1	Modes of Convergence . . . . .	35
7.2	Moment Generating Functions . . . . .	37
7.3	Central Limit Theorem . . . . .	39
7.4	The Law of Large Numbers and Inequalities . . . . .	41
<b>8</b>	<b>Estimation</b>	<b>43</b>
8.1	Estimators . . . . .	43
8.2	Maximum Likelihood Estimation . . . . .	44
8.3	Confidence Intervals . . . . .	46
<b>9</b>	<b>Hypothesis Testing</b>	<b>48</b>
9.1	Definitions . . . . .	48
9.2	Testing for a Population Mean . . . . .	49
9.2.1	Normal Distribution with Known Variance . . . . .	50
9.2.2	Normal Distribution with Unknown Variance . . . . .	50
9.3	Testing for Differences in Population Means . . . . .	50
9.3.1	Normal Distribution with Known Variances . . . . .	51
9.3.2	Normal Distribution with Unknown Variances . . . . .	51
9.4	Goodness of Fit . . . . .	51
9.4.1	Chi-square Test . . . . .	51
9.4.2	Independence using Chi-square Statistic . . . . .	52

# Chapter 1

## Review of Elementary Set Theory

$\Omega$	universal set
$\emptyset$	empty set
$A \subseteq \Omega$	subset of $\Omega$
$\overline{A}$	Complement of $A$
$ A $	cardinality of $A$
$A \cup B$	union ( $A$ or $B$ )
$A \cap B$	intersection( $A$ and $B$ )
$A = B$	both sets have exactly the same elements
$A \setminus B$	set difference (elements in $A$ that are not in $B$ )
$\{\omega\}$	a singleton with only the element $\omega$ in the set
$A \times B$	$\{(a, b)   a \in A, b \in B\}$

# Chapter 2

## Visual and Numerical Summaries

### 2.1 Visualization

**Definition 1.** The *histogram* allows us to visualize how a sample of data is distributed, say the observed values are  $\{x_1, \dots, x_n\}$ . The first step is deciding on a set of *bins* that divide the range of  $x$  into a series of intervals. A histogram then shows the *frequency* for each bin.

**Comments** Often the histogram's  $y$ -axis is normalized in some way.

- Instead of showing frequency, the height of the histogram can show *relative frequency*, the fraction of the data set contained within the bin. In this case,  $1 = \sum_{\text{bins } i} y_i$ , where  $y_i$  is the relative frequency at bin  $i$ .
- The histogram could also show the *density*, the relative frequency divided by the bin width. In this case,  $1 = \sum_{\text{bins } i} \rho_i \Delta x_i$ , where  $\rho_i$  is the density for bin  $i$  and  $\Delta x_i$  is the width of bin  $i$ .

**Definition 2.** The *empirical cumulative distribution function* of a sample of real values  $\{x_1, \dots, x_n\}$  is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x),$$

where  $I(x_i \leq x)$  is an *indicator function*, i.e. the value is 1 when  $x_i \leq x$  and 0 when  $x_i > x$ .

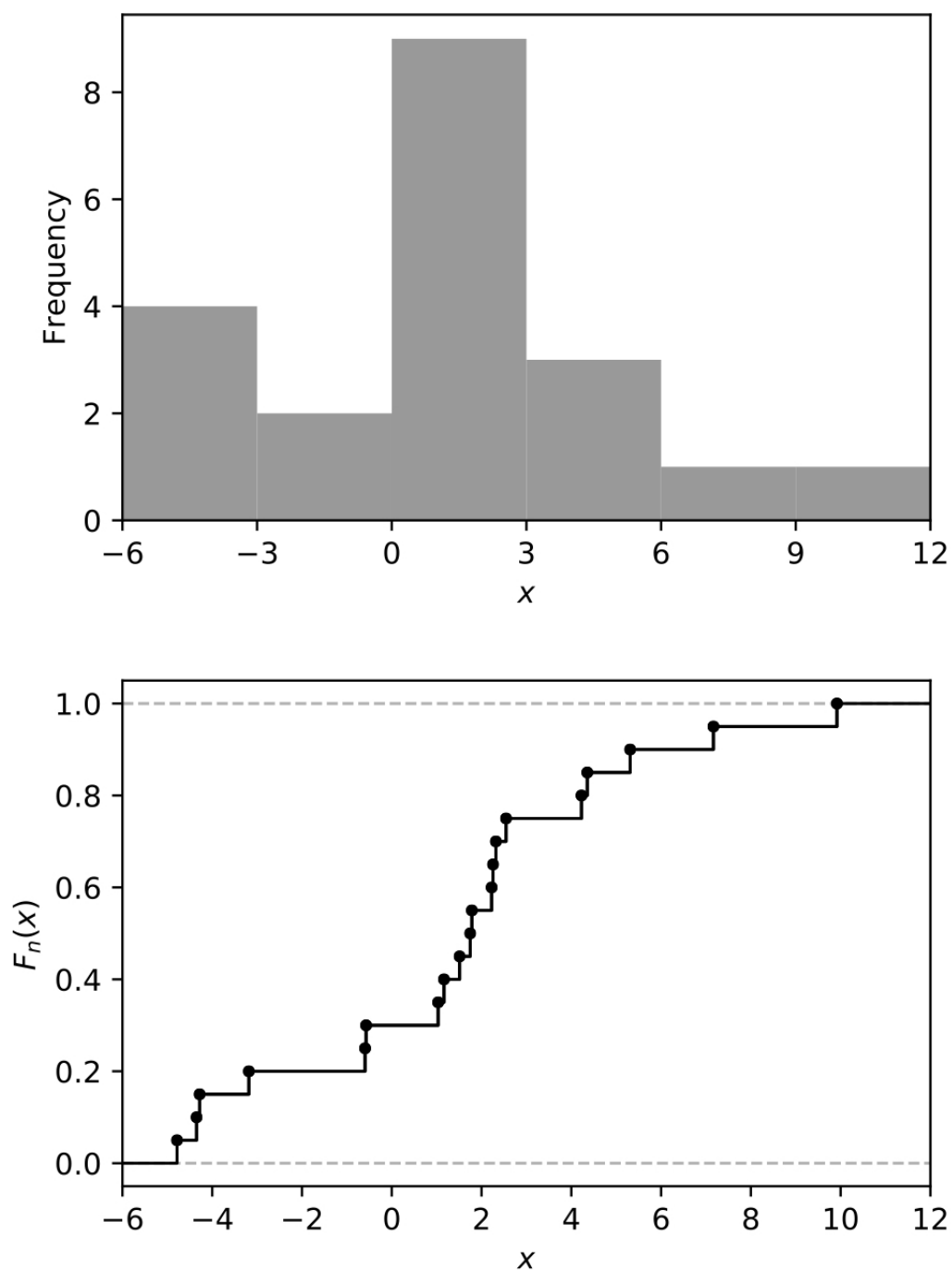


Figure 2.1: The first diagram is the histogram, and the second diagram is the empirical cdf with the same set of data

## 2.2 Summary Statistics

### 2.2.1 Measures of Location

arithmetic mean	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
geometric mean	$x_G = \left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}}$
harmonic mean	$\frac{1}{x_H} = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$
$i^{\text{th}}$ order statistic	$x_{(i)} = \text{the } i^{\text{th}} \text{ smallest value of the sample}$
median	$x_{(\frac{n+1}{2})}$
mode	$x_i \text{ which occurs most frequently in the sample}$

#### Comments

- For positive data  $\{x_1, \dots, x_n\}$ ,

$$\text{arithmetic mean} \geq \text{geometric mean} \geq \text{harmonic mean}.$$

- Arithmetic mean and geometric mean are related in the following way:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n \ln y_i = \frac{1}{n} \ln \prod_{i=1}^n y_i = \ln \left( \prod_{i=1}^n y_i \right)^{\frac{1}{n}} = \ln x_G,$$

where  $x_i = \ln y_i$ .

- For  $x_{(i)}$ , when  $i$  is not an integer, we define  $\alpha \in (0, 1)$  s.t.  $\alpha = i - \lfloor i \rfloor$ , and

$$x_{(i)} = (1 - \alpha)x_{(\lfloor i \rfloor)} + \alpha x_{(\lceil i \rceil)}.$$

### 2.2.2 Measures of Dispersion

mean square/sample variance	$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
root mean square/sample standard deviation	$s = \sqrt{s^2}$
range	$x_{(n)} - x_{(1)}$
first quartile	$x_{(\frac{1}{4}(n+1))}$
third quartile	$x_{(\frac{3}{4}(n+1))}$
interquartile range	$x_{(\frac{1}{4}(n+1))} - x_{(\frac{3}{4}(n+1))}$

#### Comments

- sample variance's different expression:

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left( \frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \left( \frac{1}{n} \sum_{i=1}^n x_i \right)^2 = \overline{x^2} - \bar{x}^2.$$

- Robustness, shown in table 2.1

Table 2.1: Robustness of different location and dispersion statistic

	Least Robust	More Robust	Most Robust
Location	$\frac{x_{(1)} + x_{(n)}}{2}$	$\bar{x}$	$x_{(\frac{n+1}{2})}$
Dispersion	$x_{(n)} - x_{(1)}$	$s$	$x_{(\frac{3}{4}(n+1))} - x_{(\frac{1}{4}(n+1))}$

### 2.2.3 Covariance, Correlation, and Skewness

covariance	$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
correlation	$r_{xy} = \frac{s_{xy}}{s_x s_y}$
skewness	$\frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$



### Comments

- covariance's different expression:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i + \frac{1}{n} \sum_{i=1}^n -x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y} = \frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \bar{y}.$$

In the random variable's context, it is

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

- Correlation gives a **scale-invariant** measurement of relatedness between  $x$  and  $y$ , since

$$|r_{xy}| \leq 1.$$

- A sample is **positively (negatively)** or **right (left) skewed** if the upper tail of the histogram of the sample is longer (shorter) than the lower tail.

### 2.2.4 Box-and-whisker plot

The diagram is based on the five-point summary (use Figure 2.2 as reference):

- Median – middle line in the box.
- 3<sup>rd</sup> and 1<sup>st</sup> Quartiles – top and bottom of the box.
- “Whiskers” – extend out as dashed lines from the box to max/min values, which are the two short horizontal lines.
- Any outliers, i.e. extreme points beyond the whiskers, are plotted individually as dots.

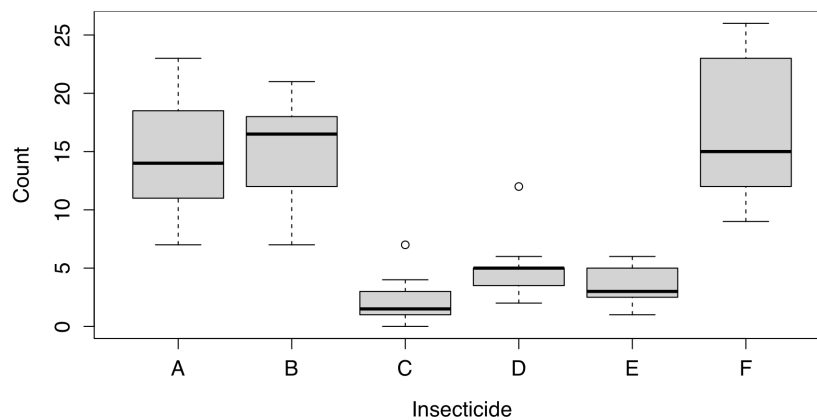


Figure 2.2: the counts of insects found in agricultural experimental units treated with six different insecticides A-F

# Chapter 3

## Probability

### 3.1 Formal Definition of Probability

#### 3.1.1 $\sigma$ -algebra

**Definition 3.**  $\mathcal{F}$ , a collection of subsets of a set  $S$ , is called a  $\sigma$ -*algebra* associated with  $S$  if:

- (a)  $S \in \mathcal{F}$ ,
- (b)  $\mathcal{F}$  is closed under complements w.r.t.  $S$ :

$$E \in \mathcal{F} \implies \overline{E} \in \mathcal{F},$$

- (c)  $\mathcal{F}$  is closed under countable unions:

$$E_1, E_2, \dots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} E_i \in \mathcal{F}.$$

**Comments** Definition 3 implies two facts.

1.  $\mathcal{F}$  must contain the empty set  $\emptyset$ .

*Proof.* Since  $S \in \mathcal{F}$ , we have  $\overline{S} = \emptyset \in \mathcal{F}$ . □

2.  $\mathcal{F}$  must be closed under countable intersections.

*Proof.* Let  $E_1, E_2, \dots \in \mathcal{F}$ . We can then imply the following:

$$\overline{E_1}, \overline{E_2}, \dots \in \mathcal{F} \Rightarrow \bigcup_i \overline{E_i} \in \mathcal{F} \Rightarrow \overline{\bigcup_i \overline{E_i}} \in \mathcal{F} \xrightarrow{\text{De Morgan's Law}} \bigcap_i E_i \in \mathcal{F}.$$

□

In short, we can take unions, intersections, and complements of members of  $\mathcal{F}$  in any combination and the result will always be a member of  $\mathcal{F}$ .

### 3.1.2 Probability Measure

**(Kolmogorov's axioms of probability) Definition 4.** A *probability measure*  $P$  is a function  $P : \mathcal{F} \mapsto \mathbb{R}$  satisfying

- (a)  $P(E) \geq 0 \forall E \in \mathcal{F}$ ,
- (b)  $P(S) = 1$ ,
- (c) If  $E_1, E_2, \dots \in \mathcal{F}$  are disjoint (i.e.  $E_i \cap E_j = \emptyset \forall i \neq j$ ) then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

The triplet  $(S, \mathcal{F}, P)$ , consisting of a set  $S$ , a  $\sigma$ -algebra  $\mathcal{F}$  of subsets of  $S$ , and a probability measure  $P$ , is called a *probability space*.

#### Comments

- The *sample space* ( $S$ ) is the set of all possible outcomes of an experiment.
- The *event space* ( $\mathcal{F}$ ) is the set of possible events, where an *event*  $E$  is a subset of the sample space,  $E \subseteq S$ . An *elementary event* is one that consist of a single element of  $S$ , i.e. a singleton.
- The probability measure ( $P$ ) has three important interpretations:
  1. **classical**: Different outcomes in the sample space  $S$  are “equally likely”,
  2. **frequentist**: the relative frequency of an event over many trials,
  3. **subjective**: a numerical measure of the degree of belief held by an individual.

**Example 5.** “A sensor can detect items within 10 cm of the sensor. The sensor is placed in a room together with an object, and the probability that the sensor makes a detection is 0.0001.”

1. **classical**: The volume within 10 cm of the sensor divided by the volume of the room is 0.0001.
  2. **frequentist**: If we repeat the experiment a lot of times, then the fraction of the experiments in which the sensor makes a detection is 0.0001.
  3. **subjective**: Someone's subjective degree of belief, measured on a numerical scale from 0 to 1, that the sensor will detect is 0.0001.
- several results that can be derived from the probability measure axioms:
    - $P(\emptyset) = 0$ .

- $P(E) \leq 1$ .
- $P(\overline{E}) = 1 - P(E)$ .
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$ .
- $P(E \cap \overline{F}) = P(E) - P(E \cap F)$ .
- If  $E \subset F$  then  $P(E) \leq P(F)$ .

## 3.2 Conditional Probability

**Definition 6.** If  $P(F) > 0$  then the *conditional probability* of  $E$  given  $F$  is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

### Comments

- Difference among the following forms:
  - $P(E|F)$  – *conditional probabilities*,
  - $P(E \cap F)$  – *joint probabilities*,
  - $P(E)$  – *marginal probabilities*.
- several results derived from the conditional probability definition:
  - $P(E|F) \geq 0$  for any event  $E$ .
  - $P(F|F) = 1$ .
  - If the events  $E_1, E_2, \dots$  are pairwise disjoint, then  $P\left(\left(\bigcup_i E_i\right) | F\right) = \sum_i P(E_i|F)$ .
- Warning: In general,  $P(E|F) \neq P(F|E)$ .

**Example 7.** A medical test for a disease  $D$  has outcomes  $+$  and  $-$ . The probabilities are

	$D$	$\overline{D}$	
$+$	0.009	0.099	0.108
$-$	0.001	0.891	0.892
	0.01	0.99	

By the definition of conditional probability, we have

$$P(+|D) = 90\%, \quad P(-|\overline{D}) = 90\%, \quad P(D|+) = \frac{0.009}{0.108} \approx 0.083.$$

The first two probabilities show that the test is fairly accurate. Sick people yield a positive 90% of the time and healthy people yield a negative 90% of the time.

### 3.3 Independence

**Definition 8.** Two events  $E$  and  $F$  are *independent* iff

$$P(E \cap F) = P(E)P(F).$$

#### Comments

- Extension: The events  $E_1, \dots, E_k$  are independent if, for every subset of events of size  $l \leq k$ , say indexed by  $\{i_1, \dots, i_l\}$ ,

$$P\left(\bigcap_{j=1}^l E_{i_j}\right) = \prod_{j=1}^l P(E_{i_j}).$$

- Independence could be either assumed or verified via the definition.
- Disjoint events with positive probability are not independent.
- From the definition of conditional probability, we can deduce that  $E$  and  $F$  are independent iff  $P(E|F) = P(E)$ .

**Definition 9.** For three events  $E_1, E_2, F$ , the pair of events  $E_1$  and  $E_2$  are said to be *conditionally independent given  $F$*  iff

$$P(E_1 \cap E_2 | F) = P(E_1 | F)P(E_2 | F).$$

which could also be written as  $E_1 \perp E_2 | F$ .

### 3.4 Bayes' Theorem

**(The Law of Total Probability) Theorem 10.** Let  $E_1, E_2, \dots$  be a partition of  $S$ , i.e.  $E_i \cap E_j = \emptyset$  for  $i \neq j$  and  $\bigcup_i E_i = S$ . Then, for any event  $F \subseteq S$ , we have

$$P(F) = \sum_i P(F|E_i)P(E_i).$$

*Proof.*  $P(F) = P(\bigcup_i F \cap E_i) = \sum_i P(F \cap E_i) = \sum_i P(F|E_i)P(E_i)$ . □

**(Bayes' Theorem) Theorem 11.** If  $P(F) > 0$  and let  $E_1, E_2, \dots$  be a partition on  $S$  s.t.  $P(E_i) > 0 \forall i$ , we have

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{P(F)} = \frac{P(F|E_i)P(E_i)}{\sum_j P(F|E_j)P(E_j)},$$

where  $P(E_i|F)$  is called the **posterior**,  $P(F|E_i)$  is called the **likelihood**,  $P(E_i)$  is called the **prior**, and  $P(F)$  is called the **evidence**.

*Proof.* Exercise! haha □

**Example 12.** A new covid-19 test is claimed to correctly identify 95% of people who are really covid-positive and 98% of people who are really covid-negative. If only 1 in a 1000 of the population are infected, what is the probability that a randomly selected person who tests positive actually has the disease?

Let  $I$  = “has a covid infection” and  $T$  = “test is positive”. We are given  $P(T|I) = 0.95$ ,  $P(\bar{T}|\bar{I}) = 0.98$ ,  $P(I) = 0.001$ . We can thus derive that

$$P(I|T) = \frac{P(T|I)P(I)}{P(T)} = \frac{P(T|I)P(I)}{P(T|I)P(I) + P(T|\bar{I})P(\bar{I})} = \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 \times 0.999} = 0.045.$$

# Chapter 4

## Discrete Random Variables

### 4.1 Random Variables

**Definition 13.** A *random variable* is a (measurable) mapping

$$X : S \mapsto \mathbb{R}$$

with the property that  $\{s \in S : X(s) \leq x\} \in \mathcal{F} \forall x \in \mathbb{R}$ . This ensures that any set  $B \subseteq \mathbb{R}$  corresponds to an event in the event space  $\mathcal{F}$ .

**Definition 14.** The image of  $S$  under  $X$  is called the *range* of the random variable

$$\mathbb{X} \equiv X(S) = \{X(s) | s \in S\} = \{x \in \mathbb{R} | \exists s \in S \text{ s.t. } X(s) = x\}.$$

So  $S$  contains all the possible outcomes of the experiment,  $\mathbb{X}$  contains all the possible outcomes of the random variable  $X$ .

**Definition 15.** The *probability distribution* of  $X$  is defined as

$$P_X = P_X(X \in B \subseteq \mathbb{R}) = P(\{s \in S : X(s) \in B\})$$

which enables us to transfer the probability measure  $P$  defined on  $\mathcal{F}$  to the real numbers in a natural way, and vice versa. For instance,

$$\begin{aligned} P_X(X = 7) &= P(\{s \in S | X(s) = 7\}), \\ P_X(a < X \leq b) &= P(\{s \in S | a < X(s) \leq b\}). \end{aligned}$$

**Example 16.** Consider counting the number of heads in a sequence of 3 coin tosses. The underlying sample space is

$$S = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}.$$

Since we are only interested in the number of heads in each sequence, we define the random variable  $X$  by

$$X(s) = \begin{cases} 0, & s = TTT, \\ 1, & s \in \{TTH, THT, HTT\}, \\ 2, & s \in \{HHT, HTH, THH\}, \\ 3, & s = HHH. \end{cases}$$

Thus, the probability of the number of heads  $X$  is less than 2 is

$$\begin{aligned} P_X(X < 2) &= P(\{s \in S : X(s) < 2\}) \\ &= P(\{TTT, TTH, THT, HTT\}) \\ &= \frac{|\{TTT, TTH, THT, HTT\}|}{|S|} \\ &= \frac{4}{8} = \frac{1}{2}. \end{aligned}$$

On a side note, the above process uses the classical interpretation on the probability measure.

**Definition 17.** The *Cumulative Distribution Function (CDF)* of a random variable  $X$  is the function  $F_X : \mathbb{R} \mapsto [0, 1]$ , defined by

$$F_X(x) = P_X(X \leq x) = P(\{s \in S : X(s) \leq x\}).$$

### Comments

- Given a right-continuous function  $F_X(x)$ , check the following to verify if it is a valid CDF:

- (i)  $0 \leq F_X(x) \leq 1 \forall x \in \mathbb{R}$ ,
- (ii) Monotonicity (non-decreasing):  $\forall x_1, x_2 \in \mathbb{R}, x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$ .
- (iii)  $F_X(-\infty) = 0, F_X(\infty) = 1$ .

- For finite intervals  $(a, b] \subseteq \mathbb{R}$ , it is easy to check that

$$P_X(a < X \leq B) = F_X(b) - F_X(a).$$

- Usually we suppress the subscript of  $P_X(\cdot)$  and just write  $P(\cdot)$  for the probability measure for the random variable, unless there is any ambiguity.

## 4.2 Discrete Random Variables



**Definition 18.** A random variable  $X$  is **discrete** if the range of  $X$ ,  $\mathbb{X}$ , is countable, that is

$$\mathbb{X} = \{x_1, x_2, \dots, x_n\} \text{ (finite)} \quad \text{or} \quad \mathbb{X} = \{x_1, x_2, \dots\} \text{ (infinite)}.$$

**Definition 19.** For a discrete random variable  $X$ , we define the **Probability Mass Function (PMF)** as

$$p_X(x) = P_X(X = x), \quad x \in \mathbb{X}.$$

For completeness, we also define

$$p_X(x) = 0, \quad x \notin \mathbb{X}.$$

so that  $p_x$  is defined for all  $x \in \mathbb{R}$ .

**Definition 20.** The **support** of a random variable  $X$  is defined as

$$\{x \in \mathbb{R} : p_X(x) > 0\},$$

which is almost always the same as the range  $\mathbb{X}$ .

### Properties of $p_X$ and $F_X$

- $p_X(x_i) \geq 0$ .
- $\sum_{x \in \mathbb{X}} p_X(x) = 1$ .
- $F_X(x) = P(X \leq x)$ ,  $x \in \mathbb{R}$ .
- Let  $X$  be a discrete random variable with range  $\mathbb{X} = \{x_1, x_2, \dots\}$ , where  $x_1 < x_2 < \dots$ . Then for any  $x \in \mathbb{R}$ , if  $x < x_1$ ,  $F_X(x) = 0$ ; otherwise

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i) \iff p_X(x_i) = F_X(x_i) - F_X(x_{i-1}), \quad i = 2, 3, \dots,$$

with  $p_X(x_1) = F_X(x_1)$ .

- $\lim_{x \rightarrow -\infty} F_X(x) = 0$ ,  $\lim_{x \rightarrow \infty} F_X(x) = 1$ .
- $F_X$  is continuous from the right on  $\mathbb{R}$ , i.e. for  $x \in \mathbb{R}$ ,  $\lim_{h \rightarrow 0^+} F_X(x+h) = F_X(x)$ .
- $F_X$  is non-decreasing, i.e.  $a < b \implies F_X(a) \leq F_X(b)$ .
- For  $a < b$ ,  $P(a < X \leq b) = F_X(b) - F_X(a)$ .

### 4.3 Functions of a Discrete Random Variable

**Definition 21.** The PMF of  $Y = g(X)$  is found by grouping all the values in the range of  $x$  that correspond to the same value of  $Y$ , i.e.

$$p_Y(y) = \sum_{x \in \mathbb{X}: g(x)=y} p_X(x).$$

### 4.4 Mean and Variance

**Definition 22.** The *expected value*, or *mean* of a discrete random variable  $X$  is defined to be

$$E_X(X) = \sum_{x \in \mathbb{X}} xp_X(x),$$

which is often written as  $E(X)$ ,  $E[X]$ , or  $\mu_X$ .

**Theorem 23.**

$$E(g(X)) = \sum_{x \in \mathbb{X}} g(x)p_X(x).$$

*Proof.* Let  $Y = g(X)$ , then

$$\begin{aligned} E(Y) &= \sum_{y \in \mathbb{Y}} yp_Y(y) \\ &= \sum_{y \in \mathbb{Y}} y \sum_{x \in \mathbb{X}: g(x)=y} p_X(x) \\ &= \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}: g(x)=y} g(x)p_X(x) \\ &= \sum_{x \in \mathbb{X}} g(x)p_X(x). \end{aligned}$$

□

**Theorem 24.** Let  $X$  be a random variable with  $p_X$ . Let  $g$  and  $h$  be real-valued functions,  $g, h : \mathbb{R} \mapsto \mathbb{R}$ , and let  $a$  and  $b$  be constants. Then

$$E(ag(X) + bh(X)) = aE(g(X)) + bE(h(X)).$$

*Proof.* Exercise!

□

**Definition 25.** Let  $X$  be a random variable. The **variance** of  $X$ , denoted by  $\sigma^2$  or  $\sigma_X^2$  or  $\text{Var}_X(X)$ , is defined by

$$\text{Var}_X(X) = E_X \left[ (X - E_X(X))^2 \right].$$

**Proposition 26.**

$$\text{Var}(X) = E(X^2) - E(X)^2.$$

*Proof.*

$$\begin{aligned} \text{LHS} &= E \left[ X^2 - 2E(X)X + E(X)^2 \right] \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= \text{RHS}. \end{aligned}$$

□

**Proposition 27.**

$$\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) \pm 2ab\text{Cov}(X, Y).$$

*Proof.* Exercise!

□

**Definition 28.** The **standard deviation** of a random variable  $X$ , written  $\text{sd}_X(X)$  or  $\sigma_X$ , is the square root of the variance,

$$\sigma_X = \sqrt{\text{Var}_X(X)}.$$

**Definition 29.** The **skewness** ( $\gamma_1$ ) of a discrete random variable  $X$  is given by

$$\gamma_1 = \frac{E_X \left[ \{X - E_X(X)\}^3 \right]}{\sigma_X^3}.$$

### Sums of Random Variables

Let  $X_1, X_2, \dots, X_n$  be  $n$  random variables, perhaps with different distributions and not necessarily independent. Let  $S_n = \sum_{i=1}^n X_i$  be the sum of those variables, and  $\frac{S_n}{n}$  be their sample average. Both  $S_n$  and  $\bar{S} = \frac{S_n}{n}$  are random variables themselves.

The mean of  $S_n$  and  $\frac{S_n}{n}$  are given by

$$E(S_n) = \sum_{i=1}^n E(X_i), \quad E\left(\bar{S}\right) = \frac{\sum_{i=1}^n E(X_i)}{n} = \mu_X.$$

If  $X_1, X_2, \dots, X_n$  are **independent**, we can calculate the variance of  $S_n$  and  $\bar{S} = \frac{S_n}{n}$  as well:

$$\text{Var}(S_n) = \sum_{i=1}^n \text{Var}(X_i), \quad \text{Var}(\bar{S}) = \frac{\sum_{i=1}^n \text{Var}(X_i)}{n^2} = \frac{\sigma_X^2}{n}.$$

## 4.5 Some important Discrete Random Variables

**Definition 30.** We say  $X$  follows a **Bernoulli Distribution** if  $X \sim \text{Bernoulli}(p)$ , where  $0 \leq p \leq 1$ , and the pmf is given by

$$p_X(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \\ 0 & \text{otherwise} \end{cases} = p^x(1-p)^{1-x}, \quad x \in \mathbb{X} = \{0, 1\}.$$

**Definition 31.** We say  $X$  follows a **Binomial Distribution** if  $X \sim \text{Binomial}(n, p)$ , where  $0 \leq p \leq 1$  and  $n \in \mathbb{Z}^+$ , and the pmf is given by

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \mathbb{X} = \{0, 1, 2, \dots, n\}.$$

**Definition 32.** We say  $X$  follows a **Geometric Distribution** if  $X \sim \text{Geometric}(p)$ , where  $0 \leq p \leq 1$ , and the pmf is given by

$$p_X(x) = p(1-p)^{x-1}, \quad x \in \mathbb{X} = \{1, 2, \dots\}.$$

Alternatively, let  $Y = X - 1$ , then  $Y \sim \text{Geometric}(p)$  with the pmf

$$p_Y(y) = p(1-p)^y, \quad y \in \mathbb{N} = \{0, 1, 2, \dots\}.$$

**Definition 33.** We say  $X$  follows a **Poisson Distribution** if  $X \sim \text{Poissons}(\lambda)$ , where  $\lambda > 0$ , and the pmf is given by

$$p_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x \in \mathbb{X} = \{0, 1, 2, \dots\}.$$

**Definition 34.** We say  $X$  follows a **Discrete Uniform Distribution** if  $X \sim \text{Uniform}(\{1, 2, \dots, n\})$ , and the pmf is given by

$$p_X(x) = \frac{1}{n}, \quad x \in \mathbb{X} = \{1, 2, \dots, n\}.$$

Table 4.1: Means and Variances of different distributions

	Mean( $\mu$ )	Variance( $\sigma^2$ )	Skewness( $\gamma_1$ )
Bernoulli	$p$	$p(1 - p)$	N.A.
Binomial	$np$	$np(1 - p)$	$\frac{1 - 2p}{\sqrt{np(1 - p)}}$
Geometric(original)	$\frac{1}{p}$	$\frac{1 - p}{p^2}$	$\frac{2 - p}{\sqrt{1 - p}}$
Geometric(alternative)	$\frac{1 - p}{p}$	$\frac{1 - p}{p^2}$	$\frac{2 - p}{\sqrt{1 - p}}$
Poisson	$\lambda$	$\lambda$	$\frac{1}{\sqrt{\lambda}}$
Uniform	$\frac{n + 1}{2}$	$\frac{n^2 - 1}{12}$	0

**Comments**

- From table 4.1, we can see that the skewness of both Geometric and Poisson Distribution is always positive.
- **Approximation of Binomial distribution as Poisson distribution.** It can be shown that for Binomial( $n, p$ ), when  $p$  is small and  $n$  is large, this distribution can be well approximated by the Poisson distribution with rate parameter  $\lambda = np$ , Poisson( $np$ ).

# Chapter 5

## Continuous Random Variable

### 5.1 Definitions

**Definition 35.** A random variable  $X$  is (absolutely) continuous if  $\exists f_X : \mathbb{R} \mapsto \mathbb{R}$  (measurable) s.t.  $f_X$  is non-negative and

$$P(X \in B) = \int_{x \in B} f_X(x) dx, \quad B \subseteq \mathbb{R},$$

and  $f_X$  is referred to as the ***Probability Density Function (PDF)*** of  $X$ .

#### Comments

- It follows that  $f_X$  is a pdf for a continuous variable  $X$  iff

$$f_X(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f_X(x) dx = 1.$$

- The pdf  $f_X(x)$  is *not* a probability. It is a probability *density*, having units of  $1/[\text{units of } X]$ . As such,

$$\forall x \in \mathbb{R}, P(X = x) = 0.$$

- Since the pdf is not itself a probability, unlike the pmf of a discrete random variable, we do *not* require  $f_X(x) \leq 1$ .

**Definition 36.** The ***Cumulative Distribution Function (CDF)***,  $F_X$ , of a continuous random variable  $X$  is defined as

$$F_X(x) = P(X \leq x), x \in \mathbb{R}.$$

**Comments**

- From now on, we implicitly assume the absolutely continuous case, then the CDF can be written as

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(x') dx', \quad x \in \mathbb{R}.$$

- For the cdf of a continuous random variable,

$$\lim_{x \rightarrow -\infty} F_X(x) = 0, \quad \lim_{x \rightarrow \infty} F_X(x) = 1.$$

- At values of  $x$  where  $F_X$  is differentiable,

$$f_X(x) = \left. \frac{d}{dt} F_X(t) \right|_{t=x} \equiv F'_X(x).$$

- For  $a < b$ ,

$$P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b) = P(a < X < b) = F_X(b) - F_X(a).$$

**5.2 Transformations**

Let  $X$  be a continuous random variable with pdf  $f_X$  and cdf  $F_X$ . Let  $Y = g(X)$  be a *transformation* (function) of  $X$  for some (measurable) function  $g : \mathbb{R} \mapsto \mathbb{R}$  s.t.  $g$  is continuous. Given  $f_X$ , how do we obtain  $f_Y$ ?

**Method 1**

1. Integrate  $f_X(x)$  to find  $F_X(x)$ .
2. Find  $F_Y(y)$  in terms of  $F_X(x)$ .
3. Differentiate  $F_Y(y)$  to get pdf  $f_Y(y)$ .

**Example 37.** Given  $f_X(x) = e^{-x}$  for  $x > 0$ . Thus,

$$F_X(x) = \int_0^x f_X(u) du = 1 - e^{-x}.$$

Let  $Y = g(X) = \log(X)$ . Then the range of  $Y$  is  $\mathbb{R}$  and

$$F_Y(y) = P(Y \leq y) = P(\log(X) \leq y) = P(X \leq e^y) = F_X(e^y).$$

Taking the derivative of the cdf gives the pdf

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(e^y) = e^y f_X(e^y) = e^y e^{-e^y}.$$

**Method 2**

1. Go to the pdf directly by matching the equation  $f_Y(y)dy = f_X(x)dx$ .

**Example 38.** Given  $f_X(x) = e^{-x}$  for  $x > 0$  and let  $Y = g(X) = \log(X)$ . Then the range of  $Y$  is  $\mathbb{R}$ . We can then obtain

$$x = g^{-1}(y) = e^y, \quad \frac{dy}{dx} = \frac{1}{x} \Rightarrow dx = |x dy| = e^y dy.$$

The absolute sign is to ensure that the product  $f_X(x_i)dx_i$  is not negative. Fitting into the equation, we have

$$f_Y(y)dy = f_X(x)dx = f_X(e^y)e^y dy.$$

We can thus obtain

$$f_Y(y) = f_X(e^y)e^y = e^y e^{-e^y}.$$

**Warning**  $g$  may not be a 1-to-1 function, e.g.  $Y = X^2$ . In this case, always draw a graph and think about the ranges of  $X$  and  $Y$ . Following the example of  $Y = X^2$ , we can derive that

$$x = \pm\sqrt{y}, \quad \frac{dy}{dx} = 2x = \pm 2\sqrt{y},$$

and then note the following

$$\begin{aligned} f_Y(y)dy &= f_X(x)dx = f_X(\sqrt{y}) \left| \frac{dy}{2\sqrt{y}} \right| + f_X(-\sqrt{y}) \left| \frac{dy}{-2\sqrt{y}} \right|, \\ \Rightarrow f_Y(y) &= \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}. \end{aligned}$$

**Example 39.**  $X \sim \text{Uniform}(-1, 3)$ , let  $Y = X^2$ . Find  $f_Y(y)$ .

Firstly, we have

$$f_X(x) = \begin{cases} \frac{1}{4}, & -1 \leq x \leq 3, \\ 0, & \text{otherwise} \end{cases}, \quad f_Y(y) = \frac{dx}{dy} f_X(x).$$

And we can also obtain that

$$x = \pm\sqrt{y}, \quad \frac{dx}{dy} = \pm \frac{1}{2\sqrt{y}}.$$

Thus when  $0 \leq y \leq 1$ :

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left( \frac{1}{4} + \frac{1}{4} \right) = \frac{1}{4\sqrt{y}}.$$

and when  $1 < y \leq 9$ :

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left( \frac{1}{4} \right) = \frac{1}{8\sqrt{y}}.$$

Finally for other values of  $y$ , we have  $f_Y(y) = 0$ .



### 5.3 Mean, Variance and Quantiles

**Definition 40.** For a continuous random variable  $X$  we define the *mean* or *expectation* of  $X$  as

$$\mu_X \text{ or } E_X(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

#### Comments

- More generally, for a (measurable) function of interest  $g : \mathbb{R} \mapsto \mathbb{R}$  we have

$$E_X(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

- Linearity of expectation:

$$E[ag(X) + b] = aE[g(X)] + b, \quad \forall a, b \in \mathbb{R}, g : \mathbb{R} \mapsto \mathbb{R}.$$

**Definition 41.** The *variance* of a continuous random variable  $X$  is given by

$$\sigma_X^2 \text{ or } \text{Var}_X(X) = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx.$$

#### Comments

- Equivalently,

$$\text{Var}_X(X) = \int_{-\infty}^{\infty} x^2 f_X(x) dx - \mu_X^2 = E(X^2) - E(X)^2.$$

- For a linear transformation  $aX + b$  we again have

$$\text{Var}(aX + b) = a^2 \text{Var}(X), \quad \forall a, b \in \mathbb{R}.$$

**Definition 42.** For a (continuous) random variable  $X$  we define the  $\alpha$ -*quantile*  $Q_X(\alpha)$ ,  $0 \leq \alpha \leq 1$  to satisfy  $P(X \leq Q_X(\alpha)) = \alpha$ ,

$$Q_X(\alpha) = F_X^{-1}(\alpha).$$

In particular, the *median* of  $X$  is  $Q_X(\frac{1}{2})$ .

## 5.4 Some Important Continuous Random Variables

**Definition 43.** We say that  $X$  follows a *continuous uniform distribution* on the interval  $(a, b)$ , where  $a < b$ , if  $X \sim U(a, b)$ , and the pdf is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

The cdf is given by

$$F_X(x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x < b, \\ 1, & x \geq b. \end{cases}$$

The distribution  $X \sim U(0, 1)$  is referred to as the *standard uniform distribution*.

**Definition 44.** We say that  $X$  follows a *exponential distribution* if  $X \sim \text{Exp}(\lambda)$ , where  $\lambda > 0$ , and the pdf is given by

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

The cdf is given by

$$F_X(x) = 1 - e^{-\lambda x}, \quad x \geq 0.$$

### Comments

- Interpretation: For  $T \sim \text{Exp}(\lambda)$ ,  $T$  can be interpreted as the time until an event occurs, where events occur at an “average rate”  $\lambda$ . The exponential distribution is the continuous version of the geometric distribution.
- “Lack of memory”: If we have already waited for a time  $t$ , what is the probability of still waiting at time  $t + s$ ?

$$\begin{aligned} P(X > s + t | X > t) &= \frac{P(X > s + t \cap X > t)}{P(X > t)} \\ &= \frac{P(X > s + t)}{P(X > t)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} \\ &= P(X > s). \end{aligned}$$

In words, the knowledge that we have waited for time  $s$  for an event tells us nothing about how much longer we will have to wait, i.e. the process has *no memory*. This is

known as the **Lack of Memory** property, and is unique to the exponential distribution amongst continuous distributions.

- Relation with Poisson distribution: If events in a random process occur according to a Poisson distribution with rate  $\lambda$ , then the time between events has an Exponential distribution with rate parameter  $\lambda$ .

*Proof.* Suppose we have some random event process such that  $\forall x > 0$ , the number of events occurring in  $[0, x]$ ,  $N_x$ , follows a Poisson distribution with rate parameter  $\lambda$ , so  $N_x \sim \text{Poisson}(\lambda x)$ . Such a process is known as a *Homogeneous Poisson process*. Let  $X$  be the time until the first event of this process arrives. Then we notice that

$$P(X > x) = P(N_x = 0) = \frac{(\lambda x)^0 e^{-\lambda x}}{0!} = e^{-\lambda x}.$$

Hence  $X \sim \text{Exp}(\lambda)$ . This argument applies for all subsequent inter-arrival times.  $\square$

**Definition 45.** We say that  $X$  follows a **Gaussian** or **normal distribution** if  $X \sim N(\mu, \sigma^2)$ , where  $\mu \in \mathbb{R}$ ,  $\sigma > 0$ , and the pdf is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}.$$

The cdf is given by

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp \left\{ -\frac{(t - \mu)^2}{2\sigma^2} \right\} dt.$$

### Comments

- If  $X \sim N(0, 1)$ , then  $X$  has a **standard** or **unit normal distribution**. The pdf of the standard normal distribution is written as  $\phi(x)$ , and the cdf is written as  $\Phi(x)$ .
- If  $Y \sim N(0, 1)$ , and  $X = \sigma Y + \mu$ , (or equivalently  $Y = \frac{X - \mu}{\sigma}$ ) then  $X \sim N(\mu, \sigma^2)$ . We can then write the cdf of  $X$  in terms of  $\Phi$ ,

$$F_X(x) = P(X \leq x) = P\left(Y \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

- Because the standard normal pdf  $\phi$  is *symmetric* about 0, i.e.  $\phi(-z) = \phi(z)$  for  $z \in \mathbb{R}$ , for the cdf we have

$$\Phi(z) = 1 - \Phi(-z).$$

Table 5.1: Means and Variances of different continuous distributions

	Mean( $\mu$ )	Variance( $\sigma^2$ )
Uniform	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	$\mu$	$\sigma^2$

# Chapter 6

## Jointly Distributed Random Variables

### 6.1 Definitions

**Definition 46.** Given a pair of random variables,  $X$  and  $Y$ , we define the *joint probability distribution*  $P_{XY}$  as follows:

$$\begin{aligned} P_{XY}(B_X, B_Y) &= P(X^{-1}(B_X) \cap Y^{-1}(B_Y)) \\ &= P\left(\{s \in S : X(s) \in B_X, Y(s) \in B_Y\}\right), \quad B_X, B_Y \subseteq \mathbb{R}. \end{aligned}$$

More generally, for some  $B_{XY} \subseteq \mathbb{R}^2$ , find the collection of sample space elements (i.e. the event)

$$S_{XY} = \left\{s \in S : (X(s), Y(s)) \in B_{XY}\right\},$$

and define

$$P_{XY}(B_{XY}) = P(S_{XY}).$$

**Definition 47.** Given a pair of random variables,  $X$  and  $Y$ , the *joint cumulative distribution function* is defined as

$$F_{XY}(x, y) = P_{XY}(X \leq x, Y \leq y), \quad x, y \in \mathbb{R}.$$

#### Comments

- The marginal cdfs of  $X$  and  $Y$  can be recovered by

$$F_X(x) = F_{XY}(x, \infty), \quad F_Y(y) = F_{XY}(\infty, y), \quad x, y \in \mathbb{R}.$$

- $\forall x, y \in \mathbb{R}$ ,

$$0 \leq F_{XY}(x, y) \leq 1,$$

$$F_{XY}(x, -\infty) = 0, \quad F_{XY}(-\infty, y) = 0, \quad F_{XY}(\infty, \infty) = 1.$$

- Monotonicity:  $\forall x, y \in \mathbb{R}$ , we have

$$x_1 < x_2 \Rightarrow F_{XY}(x_1, y) \leq F_{XY}(x_2, y), \quad y_1 < y_2 \Rightarrow F_{XY}(x, y_1) \leq F_{XY}(x, y_2).$$

- By noting that  $P_{XY}(x_1 < X \leq x_2, Y \leq y) = F_{XY}(x_2, y) - F_{XY}(x_1, y)$ , we can also obtain that

$$P_{XY}(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1).$$

**Definition 48.** If  $X$  and  $Y$  are both discrete random variables, then we can define the **joint probability mass function** as

$$p_{XY}(x, y) = P_{XY}(X = x, Y = y), \quad x, y \in \mathbb{R}.$$

### Comments

- We can recover the marginal pmfs  $p_X$  and  $p_Y$ , by the law of total probability,  $\forall x, y \in \mathbb{R}$ ,

$$p_X(x) = \sum_{y \in \mathbb{Y}} p_{XY}(x, y), \quad p_Y(y) = \sum_{x \in \mathbb{X}} p_{XY}(x, y).$$

- For  $p_{XY}$  to be a valid pmf, we need to make sure the following conditions hold:

$$0 \leq p_{XY}(x, y) \leq 1, \forall x, y \in \mathbb{R} \quad \text{and} \quad \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} p_{XY}(x, y) = 1.$$

**Definition 49.** If  $\exists f_{XY} : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$  s.t.  $\forall B_{XY} \subseteq \mathbb{R} \times \mathbb{R}$ ,

$$P_{XY}(B_{XY}) = \int_{(x,y) \in B_{XY}} f_{XY}(x, y) dx dy,$$

then we say  $X$  and  $Y$  are **jointly continuous** and we refer to  $f_{XY}$  as the **joint probability density function** of  $X$  and  $Y$ . In this case we have

$$F_{XY}(x, y) = \int_{t=-\infty}^y \int_{s=-\infty}^x f_{XY}(s, t) ds dt, \quad x, y \in \mathbb{R}.$$

By the fundamental theorem of calculus we can identify the joint pdf of  $X$  and  $Y$  as

$$f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y).$$

**Comments**

- We can recover the marginal densities  $f_X$  as

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{d}{dx}F_{XY}(x, \infty) = \frac{d}{dx} \int_{y=-\infty}^{\infty} \int_{s=-\infty}^x f_{XY}(s, y) ds dy,$$

and by the fundamental theorem of calculus, we obtain (similarly for  $f_Y$ )

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{XY}(x, y) dy, \quad f_Y(y) = \int_{x=-\infty}^{\infty} f_{XY}(x, y) dx.$$

- For  $f_{XY}$  to be a valid pdf, we need to make sure the following conditions hold:

$$f_{XY}(x, y) \geq 0, \forall x, y \in \mathbb{R} \quad \text{and} \quad \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} f_{XY}(x, y) dx dy = 1.$$

**Example 50.** Suppose continuous random variables  $(X, Y) \in \mathbb{R}^2$  have joint pdf

$$f(x, y) = \begin{cases} 1, & |x| + |y| < \frac{1}{\sqrt{2}} \\ 0, & \text{otherwise.} \end{cases}$$

Determine the marginal pdfs of  $X$  and  $Y$ .

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{|x| - \frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}} - |x|} dy = \sqrt{2} - 2|x|.$$

Similarly, we can obtain  $f_Y(y) = \sqrt{2} - 2|y|$ .

## 6.2 Independence, Conditional Probability, Expectation

**Definition 51.** Two continuous random variables  $X$  and  $Y$  are *independent* iff

$$f_{XY}(x, y) = f_X(x)f_Y(y), \quad \forall x, y \in \mathbb{R}.$$

**Example 52.** Suppose that the lifetime,  $X$ , and brightness,  $Y$ , of a light bulb are modelled as continuous random variables. Let their joint pdf be given by

$$f(x, y) = \lambda_1 \lambda_2 e^{-\lambda_1 x - \lambda_2 y}, \quad x, y > 0.$$

Are lifetime and brightness independent?

$$f_X(x) = \int_0^{\infty} \lambda_1 \lambda_2 e^{-\lambda_1 x - \lambda_2 y} dy$$

$$\begin{aligned}
 &= \lambda_1 e^{-\lambda_1 x} \int_0^\infty \lambda_2 e^{-\lambda_2 y} dy \\
 &= \lambda_1 e^{-\lambda_1 x} \left[ -e^{-\lambda_2 y} \right]_0^\infty \\
 &= \lambda_1 e^{-\lambda_1 x}.
 \end{aligned}$$

Similarly, we have  $f_Y(y) = \lambda_2 e^{-\lambda_2 y}$ . Thus we obtain that  $f_X(x)f_Y(y) = f_{XY}(x, y)$ , indicating that the lifetime and brightness are independent.

**Definition 53.** For two random variables  $X$  and  $Y$ , we define the *conditional probability distribution*  $P_{Y|X}$  by

$$P_{Y|X}(B_Y|B_X) = \frac{P_{XY}(B_X, B_Y)}{P_X(B_X)}, \quad B_X, B_Y \subseteq \mathbb{R}.$$

**Definition 54.** For random variables  $X$  and  $Y$ , we define the *conditional probability density function*  $f_{Y|X}$  by

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}, \quad x, y \in \mathbb{R}.$$

### Comments

The random variables  $X$  and  $Y$  are independent

$$\begin{aligned}
 &\iff P_{Y|X}(B_Y|B_X) = P_Y(B_Y), & \forall B_X, B_Y \subseteq \mathbb{R}, \\
 &\iff f_{Y|X}(y|x) = f_Y(y), & \forall x, y \in \mathbb{R}.
 \end{aligned}$$

**Definition 55.** If  $X$  and  $Y$  are discrete, we define  $E(g(X, Y))$  by

$$E(g(X, Y)) = \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} g(x, y) p_{XY}(x, y).$$

If  $X$  and  $Y$  are jointly continuous, we define  $E(g(X, Y))$  by

$$E(g(X, Y)) = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy.$$

### Comments

- Expectation is always linear:

$$E_{XY}[g_1(X, Y) + g_2(X, Y)] = E_{XY}[g_1(X, Y)] + E_{XY}[g_2(X, Y)].$$



- If  $g(X, Y) = g_1(X)g_2(Y)$  and  $X$  and  $Y$  are **independent**,

$$E_{XY}[g_1(X)g_2(Y)] = E_X[g_1(X)]E_Y[g_2(Y)].$$

In particular, considering  $g(X, Y) = XY$  for independent  $X$  and  $Y$ ,

$$E_{XY}(XY) = E_X(X)E_Y(Y).$$

**Warning!** In general  $E_{XY}(XY) \neq E_X(X)E_Y(Y)$ .

**Definition 56.** If  $X$  and  $Y$  are discrete, the **conditional expectation** of  $Y$  given  $X = x$  is

$$E_{Y|X}(Y|X = x) = \sum_{y \in \mathbb{Y}} y p(y|x).$$

Similarly for the case when  $X$  and  $Y$  are continuous, we have

$$E_{Y|X}(Y|X = x) = \int_{y=-\infty}^{\infty} y f(y|x) dy.$$

**Definition 57.** We define the **correlation** of  $X$  and  $Y$  by

$$\rho_{XY} = \text{Cor}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

## 6.3 Multivariate Transformations

**(Convolution Theorem) Theorem 58.** If  $X$  and  $Y$  are *independent* random variables and  $Z = X + Y$ , then

$$p_Z(z) \text{ or } f_Z(z) = \begin{cases} \sum_{x \in \mathbb{X}} p_X(x) p_Y(z - x) & (\text{discrete case}), \\ \int_{\mathbb{R}} f_X(\omega) f_Y(z - \omega) d\omega & (\text{continuous case}). \end{cases}$$

**Example 59.** Supposed  $X \sim N(0, \sigma^2)$  and  $Y \sim N(0, 1)$  with  $X$  and  $Y$  independent. Let  $Z = X + Y$  and derive the pdf of  $Z$ .

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-x)^2}{2}} dx \\ &\quad \vdots \\ &= \frac{1}{\sqrt{2\pi(1+\sigma^2)}} e^{-\frac{z^2}{2(1+\sigma^2)}}. \\ &\implies Z \sim N(0, 1 + \sigma^2). \end{aligned}$$

**Theorem 60.** If  $X \sim N(\mu_X, \sigma_X^2)$  and  $Y \sim N(\mu_Y, \sigma_Y^2)$  with  $X$  and  $Y$  independent, then

$$Z = X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

**Theorem 61.** Suppose  $(X, Y)$  is a bivariate random variable and let  $U = g_1(X, Y)$  and  $V = g_2(X, Y)$ . Then for any  $B \subseteq \mathbb{R}^2$ ,

$$P((U, V) \in B) = P((X, Y) \in A),$$

where  $A = \{(x, y) : (g_1(x, y), g_2(x, y)) \in B\}$ . This can be generally divided into two cases to consider:

1. If  $(X, Y)$  is *discrete*: Let  $A(u, v) = \{(x, y) \in (\mathbb{X}, \mathbb{Y}) : (g_1(x, y), g_2(x, y)) = (u, v)\}$ , then

$$p_{UV}(u, v) = P(U = u, V = v) = P((X, Y) \in A(u, v)) = \sum_{\substack{(x, y): \\ g_1(x, y) = u, \\ g_2(x, y) = v}} p_{XY}(x, y).$$

2. If  $(X, Y)$  is *continuous*: We define the **Jacobian determinant**  $|J|$  s.t.  $dx dy = |J| du dv$ , where

$$|J| = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| = \left| \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} \right| = \left| \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u} \right|.$$

Then

$$f_{UV}(u, v) = f_{XY}(x, y) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|.$$

## 6.4 Gamma and Beta Distributions

**Definition 62.** The **Gamma function** is defined by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt, \quad \alpha > 0,$$

and then we say  $X$  follows the **Gamma Distribution** if  $X \sim \text{Gamma}(\alpha, \beta)$ , where  $\alpha, \beta > 0$ , and we have

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}, \quad x \in (0, \infty).$$

**Comments**

- $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$  for  $\alpha > 1$ .
- $\Gamma(1) = \int_0^\infty e^{-t} dt = 1$ .
- $\Gamma(n) = (n - 1)!$  for  $n \in \mathbb{Z}^+$ .
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ .

**Theorem 63.** If  $X \sim \text{Gamma}(\lambda, \theta)$  and  $Y \sim \text{Gamma}(\xi, \theta)$  with  $X$  and  $Y$  independent, then  $Z = X + Y \sim \text{Gamma}(\lambda + \xi, \theta)$ .

**Definition 64.** The **Beta function** is defined by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

We say  $X$  follows the **Beta Distribution** if  $X \sim \text{Beta}(\alpha, \beta)$ , where  $\alpha, \beta > 0$ , and we have

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < x < 1,$$

# Chapter 7

## Convergence Concepts and Theorems

### 7.1 Modes of Convergence

**Definition 65.** Let  $X_1, X_2, \dots, X_n, X$  be random variables. Higher order of strength implies the lower ones. In decreasing order of strength, we have

1.  $X_n$  *converges almost surely* to  $X$  if

$$P(\lim_{n \rightarrow \infty} X_n = X) = 1 \quad \text{or} \quad X_n \rightarrow_{\text{as}} X.$$

2.  $X_n$  *converges in probability* to  $X$  if

$$\forall \epsilon > 0, \lim_{n \rightarrow \infty} P(|X_n - X| > \epsilon) = 0 \quad \text{or} \quad X_n \rightarrow_{\text{p}} X.$$

3.  $X_n$  *converges in distribution (converges weakly)* to  $X$  with cdf  $F_X$  if

$$\lim_{n \rightarrow \infty} P(X_n < x) = F_X(x) \quad \text{or} \quad X_n \rightarrow_{\text{d}} X.$$

at all continuity points  $x$  of  $F_X(x)$ .

4. If  $X_n \rightarrow_{\text{d}} X$  and  $P(X = c) = 1$  for some  $c$ , we say the limiting distribution of  $X_n$  is *degenerate* at  $c$  and write  $X_n \rightarrow_{\text{d}} c$ .

#### Comments

- *Convergence in distribution* only requires the cdf of  $X_n$ 's converges to the cdf of  $X$  as  $n \rightarrow \infty$ . It does not require any dependence between the  $X_n$ 's and  $X$ , i.e. it doesn't tell anything about whether the value of  $X_n$  will be close to  $X$  for a single run of the experiment. Thus it is in some sense the weakest type of convergence. Convergence in probability says that for large enough  $n$ , *for each run of the experiment*, there is a high probability that the two values,  $X_n$  and  $X$ , will be close together.
- $X_n \rightarrow_{\text{d}} c \iff X_n \rightarrow_{\text{p}} c$  for some  $c$ .

- (Slutsky's Theorem) If  $X_n \rightarrow_d X$  and  $Y_n \rightarrow_d c$ , then

$$X_n Y_n \rightarrow_d cX \quad \text{and} \quad X_n + Y_n \rightarrow_d X + c.$$

**Example 66.** Let  $X_1, X_2, X_3, \dots$  be a sequence of random variables s.t.

$$F_{X_n}(x) = \begin{cases} 1 - (1 - \frac{1}{n})^{nx} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Show that  $X_n \rightarrow_d \text{Exponential}(1)$ .

Let  $X \sim \text{Exponential}(1)$ . For  $x \leq 0$ , we have

$$F_{X_n}(x) = F_X(x) = 0, \quad \text{for } n = 1, 2, 3, \dots$$

For  $x > 0$ , we have

$$\begin{aligned} \lim_{n \rightarrow \infty} F_{X_n}(x) &= \lim_{n \rightarrow \infty} \left( 1 - \left( 1 - \frac{1}{n} \right)^{nx} \right) \\ &= 1 - \lim_{n \rightarrow \infty} \left( 1 - \frac{1}{n} \right)^{nx} \\ &= 1 - e^{-x} \\ &= F_X(x). \end{aligned}$$

Thus, we conclude that  $X_n \rightarrow_d X$ .

**Example 67.** Let  $X$  be a random variable, and  $X_n = X + Y_n$ , where

$$E(Y_n) = \frac{1}{n} \quad \text{and} \quad (Y_n) = \frac{\sigma^2}{n},$$

where  $\sigma > 0$  is a constant. Show that  $X_n \rightarrow_p X$ .

By the triangle inequality,  $\forall a, b \in \mathbb{R}, |a + b| \leq |a| + |b|$ . Choosing  $a = Y_n - E(Y_n)$  and  $b = E(Y_n)$ , we obtain

$$|Y_n| \leq |Y_n - E(Y_n)| + \frac{1}{n}.$$

Now for any  $\epsilon > 0$ , we have

$$\begin{aligned} P(|X_n - X| \geq \epsilon) &= P(|Y_n| \geq \epsilon) \\ &\leq P(|Y_n - E(Y_n)| + \frac{1}{n} \geq \epsilon) \\ &= P(|Y_n - E(Y_n)| \geq \epsilon - \frac{1}{n}) \\ &\leq \frac{(Y_n)}{(\epsilon - \frac{1}{n})^2} \quad (\text{by Chebyshev's inequality}) \\ &= \frac{\sigma^2}{n(\epsilon - \frac{1}{n})^2} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Therefore, we conclude that  $X_n \rightarrow_p X$ .

## 7.2 Moment Generating Functions

**Definition 68.** The *moment generating function (MGF)* of the random variable  $X$  is

$$M_X(t) = E\left(e^{tX}\right),$$

provided the expectation exists in some neighborhood of zero, i.e. the expectation exists  $\forall |t| < \epsilon$  for some  $\epsilon > 0$ .

**Theorem 69.** If  $X$  has an MGF, then

$$E(X^n) = M_X^{(n)}(0) = \left. \frac{d^n}{dt^n} M_X(t) \right|_{t=0}.$$

*Proof.* When  $n = 1$ ,

$$\frac{d}{dt} M_X(t) = \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x) dx = \int_{-\infty}^{\infty} x e^{tx} f_X(x) dx = E\left[X e^{tX}\right].$$

When  $t = 0$ ,  $M_X^{(1)}(0) = E(X)$ . It is similar for  $n > 1$ . Or we could also observe the different moments using Taylor expansions at  $t = 0$ , that

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \dots$$

leading to

$$E\left[e^{tX}\right] = 1 + t \underbrace{E(X)}_{1^{\text{st}} \text{ moment}} + \frac{t^2}{2!} \underbrace{E(X^2)}_{2^{\text{nd}} \text{ moment}} + \frac{t^3}{3!} \underbrace{E(X^3)}_{3^{\text{rd}} \text{ moment}} + \dots$$

□

**Example 70.** Suppose  $X \sim N(0, 1)$ . Derive the MGF and the first four moments of  $X$ .

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = e^{\frac{1}{2}t^2}, \\ M_X^{(1)}(t) &= \frac{d}{dt} \left( e^{\frac{1}{2}t^2} \right) = te^{\frac{1}{2}t^2}, \\ M_X^{(2)}(t) &= \frac{d}{dt} \left( te^{\frac{1}{2}t^2} \right) = (t^2 + 1)e^{\frac{1}{2}t^2}, \\ M_X^{(3)}(t) &= t(t^2 + 3)e^{\frac{1}{2}t^2}, \\ M_X^{(4)}(t) &= (t^4 + 6t^2 + 3)e^{\frac{1}{2}t^2}, \\ &\vdots \end{aligned}$$

**Theorem 71.**

$$M_{aX+b}(t) = e^{bt} M_X(at).$$

*Proof.*

$$M_{aX+b}(t) = E \left[ e^{t(aX+b)} \right] = E \left[ e^{atX} e^{bt} \right] = E \left[ e^{atX} \right] e^{bt} = e^{bt} M_X(at).$$

□

**Example 72.** Suppose  $Z \sim N(0, 1)$  and  $X = \mu + Z$ . Then  $X \sim N(\mu, \sigma^2)$ .

$$M_X(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}} = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

**Theorem 73.** Let  $X_1, \dots, X_n$  be a sequence of *independent* random variables with MGFs  $M_{X_1}(t), \dots, M_{X_n}(t)$ , and let  $Z = X_1 + \dots + X_n$ . Then

$$M_Z(t) = \prod_{i=1}^n M_{X_i}(t).$$

*Proof.*

$$M_Z(t) = E \left[ e^{t(X_1 + \dots + X_n)} \right] = E \left[ \prod_{i=1}^n e^{X_i t} \right] = \prod_{i=1}^n E \left[ e^{X_i t} \right] = \prod_{i=1}^n M_{X_i}(t).$$

□

**Example 74.** Suppose  $X_i \sim \text{Gamma}(\alpha_i, \beta)$  for  $i = 1, \dots, n$  are i.i.d., and  $S = \sum_{i=1}^n X_i$ . Then

$$M_{X_i}(t) = \left( \frac{\beta}{\beta - t} \right)^{\alpha_i} \quad \text{so that} \quad M_S(t) = \left( \frac{\beta}{\beta - t} \right)^{\sum_{i=1}^n \alpha_i}$$

and so  $S \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$ .

**Theorem 75.** Let  $X_1, \dots, X_n$  be i.i.d. random variables, each with MGF,  $M_X(t)$ . Then

$$M_{\bar{X}}(t) = \left[ M_X\left(\frac{t}{n}\right) \right]^n.$$

*Proof.* Write  $\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n$ , and then apply Theorems 71 and 73. □

**Theorem 76.** If  $M_X(t)$  exists (in a neighborhood of zero), then for  $r = 0, 1, 2, \dots$

- (i)  $M_X^{(r)}(t)$  exists near zero, and
- (ii)  $E(|X^r|) < \infty$ .

**(Characterization) Theorem 77.** If the MGFs of  $X$  and  $Y$  exist and  $M_X(t) = M_Y(t)$  in a neighborhood of zero, then

$$F_X(u) = F_Y(u) \quad \forall u,$$

i.e. MGFs characterize distributions of random variables.

**Example 78.** Refer to the Gamma distribution example, which is Example 74.

**Theorem 79.** Let  $X_1, X_2, \dots$  be a countable sequence of random variables with MGFs  $M_{X_1}(t), M_{X_2}(t), \dots$  so that

$$\lim_{i \rightarrow \infty} M_{X_i}(t) = M_X(t) \quad \text{for } t \text{ in a neighborhood of zero,}$$

where  $M_X(t)$  is an MGF. Then there is a unique cdf,  $F_X$ , for which

$$\lim_{i \rightarrow \infty} F_{X_i}(x) = F_X(x) \quad \text{for all } x \text{ where } F_X(x) \text{ is continuous,}$$

that is,  $X_n \rightarrow_d X$ . The moments of  $F_X(x)$  are determined by  $M_X(t)$ .

### Comments

- Proofs of Theorems 76, 77, and 79 follow from the properties of Laplace transforms, which are beyond the scope of the course.
- We can now show convergence in distribution by showing convergence of MGF in a neighborhood of zero.
- Convergence of MGF is a sufficient, but *not* necessary condition for convergence in distribution. (The MGF may not exist in a neighborhood of zero.)
- A more general strategy involves characteristic functions,  $\phi_X(t) = E(e^{itX})$ .
- Moments alone do not characterize distributions. That is, there exists  $X$  and  $Y$  s.t.  $E(X^r) = E(Y^r)$  for  $r = 0, 1, 2, \dots$ , but  $F_X \neq F_Y$ .
  - However, if  $X$  and  $Y$  have finite support, and all moments exist, then  $F_X(u) = F_Y(u)$  for all  $u$  iff  $E(X^r) = E(Y^r)$  for  $r = 0, 1, 2, \dots$

## 7.3 Central Limit Theorem



**(Central Limit Theorem) Theorem 80.** Let  $X_1, X_2, \dots, X_n$  be a sequence of i.i.d. random variables with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2 < \infty$ , then

$$\lim_{n \rightarrow \infty} \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \quad \text{or} \quad \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow_d N(0, 1).$$

Equivalently, when  $n$  is large we have approximately

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{or} \quad \sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2).$$

*Proof.* By assumption,  $M_X(t)$  exists in a neighborhood of zero s.t.  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2$  exists. Let  $Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ . It is sufficient to show that

$$\lim_{n \rightarrow \infty} M_{Z_n}(t) = M_Z(t) = e^{\frac{t^2}{2}},$$

where  $Z \sim N(0, 1)$ . Re-write  $Z_n$  as

$$Z_n = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \left( \frac{X_1 - \mu}{\sigma} + \frac{X_2 - \mu}{\sigma} + \dots + \frac{X_n - \mu}{\sigma} \right).$$

Define  $W_i = \frac{X_i - \mu}{\sigma}$  s.t.  $E(W_i) = 0$ ,  $\text{Var}(W_i) = 1$ , and we can rewrite  $Z_n$  as

$$Z_n = \frac{1}{\sqrt{n}} (W_1 + W_2 + \dots + W_n).$$

Therefore,

$$\begin{aligned} M_{W_i}(t) &= E\left(e^{tW_i}\right) = M_{W_i}(0) + M_{W_i}^{(1)}(0)t + \frac{M_{W_i}^{(2)}(0)}{2!}t^2 + O(t^3) \\ &= 1 + (0)t + \frac{t^2}{2} + O(t^3) \\ &= 1 + \frac{t^2}{2} + O(t^3). \end{aligned}$$

Thus,

$$\begin{aligned} M_{\frac{W_i}{\sqrt{n}}}(t) &= M_{W_i}\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + O\left(\left(\frac{t}{\sqrt{n}}\right)^3\right). \\ \Rightarrow M_{Z_n}(t) &= \prod_{i=1}^n M_{\frac{W_i}{\sqrt{n}}}(t) = \left(1 + \frac{t^2}{2n} + O\left(\left(\frac{t}{\sqrt{n}}\right)^3\right)\right)^n. \end{aligned}$$

By taking  $\log[M_{Z_n}(t)]$ , and the Taylor expansion

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots = x + O(x^2),$$

we have

$$\log[M_{Z_n}(t)] = n \left[ \frac{t^2}{2n} + O\left(\left(\frac{t}{\sqrt{n}}\right)^3\right) + O\left(\frac{t^4}{n^2}\right) \right] = \frac{t^2}{2} + O\left(\frac{t^3}{n^{\frac{1}{2}}}\right) + O\left(\frac{t^4}{n}\right) \rightarrow \frac{t^2}{2}$$

as  $n \rightarrow \infty$ . Thus,

$$\lim_{n \rightarrow \infty} M_{Z_n}(t) = e^{\frac{t^2}{2}} = M_Z(t), \quad Z \sim N(0, 1),$$

i.e.

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow_d N(0, 1).$$

□

## 7.4 The Law of Large Numbers and Inequalities

**Theorem 81.** Let  $X$  be a random variable and  $g(\cdot)$  be a non-negative function. Then  $\forall r > 0$ ,

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r}.$$

*Proof.*

$$\begin{aligned} E(g(X)) &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\ &\geq \int_{\{x: g(x) \geq r\}} g(x) f_X(x) dx \\ &\geq \int_{\{x: g(x) \geq r\}} r f_X(x) dx \\ &= r \int_{\{x: g(x) \geq r\}} f_X(x) dx = r P(g(X) \geq r). \end{aligned}$$

□

**Theorem 82.** Suppose  $X_1, X_2, \dots$  is a sequence of i.i.d. random variables with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2 < \infty$ . If  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ , then  $\bar{X}_n \rightarrow_p \mu$ , i.e.

$$\forall \epsilon > 0, \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1.$$

*Proof.* Apply Chebyshev's inequality to random variable  $\bar{X}_n$ . Let  $g(x) = (x - \mu)^2$  s.t.  $E[g(\bar{X})] = E[(\bar{X} - \mu)^2] = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ . Therefore,

$$P(|\bar{X}_n - \mu| \geq \epsilon) = P((\bar{X}_n - \mu)^2 \geq \epsilon^2) \leq \frac{E[(\bar{X}_n - \mu)^2]}{\epsilon^2} = \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \rightarrow 0$$

as  $n \rightarrow \infty$ , therefore  $P(|\bar{X}_n - \mu| < \epsilon) \rightarrow 1$ .

□

**(Jensen's Inequality) Theorem 83.**

- If  $g(x)$  is a *convex* function ( $g'' \geq 0$ ), then  $E[g(X)] \geq g[E(X)]$ .
- If  $g(x)$  is a *concave* function ( $g'' \leq 0$ ), then  $E[g(X)] \leq g[E(X)]$ .

**Example 84.** Suppose  $X \sim \text{Binomial}(n, p)$  and consider the estimator  $\hat{P} = \frac{X}{n}$  of  $p$ . We can show that  $\hat{P}$  is an unbiased estimator of  $p$  by

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p.$$

Suppose we want an estimator of the odds ratio  $\xi = \frac{p}{1-p} = g(p)$ .

$$g'(p) = \frac{1-p+p}{(1-p)^2} = \frac{1}{(1-p)^2},$$

$$g''(p) = \frac{2}{(1-p)^3} > 0 \quad \text{if } p \in (0, 1).$$

Consider the estimator  $\Xi = \frac{\hat{P}}{1-\hat{P}}$  of  $\xi$ ,

$$E(\Xi) = E[g(\hat{P})] \geq g[E(\hat{P})] = g(p) = \xi,$$

implying that  $\Xi$  is biased.

In general if we have an unbiased estimator  $\hat{\theta}$  of parameter  $\theta$  and want to estimate some function of the parameter  $\phi = g(\theta)$  using the estimator  $\hat{\phi} = g(\hat{\theta})$  it is important to realize that

$$E(\hat{\phi}) = E[g(\hat{\theta})] \neq g[E(\hat{\theta})] = g(\theta) = \phi,$$

i.e. unbiasedness is *not* necessarily invariant to transformation.

# Chapter 8

## Estimation

### 8.1 Estimators

**Definition 85.**

- A **statistic** is a function  $T = T(X_1, X_2, \dots, X_n) = T(\mathbf{X})$ , and *is* itself a random variable.
- If a statistic  $T(\mathbf{X})$  is to be used to approximate parameters of the distribution  $P_{\mathbf{X}|\theta}(\cdot)$ , we say that  $T$  is an **estimator** for those parameters, and we call the actual realized value of the estimator for a particular data sample,  $t(\mathbf{x})$ , an **estimate**.
- A **point estimate** is a statistic estimating a single parameter or characteristic of a distribution.

**Definition 86.** We define the **bias** of an estimator  $T$  for a parameter  $g(\theta)$  as

$$\text{bias}_\theta(T) = E_\theta(T) - g(\theta).$$

If the estimator has zero bias, we say the estimator is **unbiased**.

**Definition 87.** We define the **bias-corrected sample variance** as

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

which is then always an unbiased estimator of the population variance  $\sigma^2$ .

**Definition 88.** Suppose we have two unbiased estimators for a parameter  $\theta$ , which we call  $\hat{\Theta}(\mathbf{X})$  and  $\hat{\Psi}(\mathbf{X})$ . We say  $\hat{\Theta}$  is *more efficient* than  $\hat{\Psi}$  if

1.  $\forall \theta, \text{Var}_{\theta}(\hat{\Theta}) \leq \text{Var}_{\theta}(\hat{\Psi})$ ,
2.  $\exists \theta$  s.t.  $\text{Var}_{\theta}(\hat{\Theta}) < \text{Var}_{\theta}(\hat{\Psi})$ .

If  $\hat{\Theta}$  is more efficient than any other possible estimator, we say  $\hat{\Theta}$  is *efficient*.

**Definition 89.** We say that  $\hat{\Theta}$  is a *consistent* estimator for the parameter  $\theta$  if  $\hat{\Theta}$  converges in probability to  $\theta$ , i.e.

$$\forall \epsilon > 0, P(|\hat{\Theta} - \theta| > \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

This is hard to demonstrate, but if  $\hat{\Theta}$  is unbiased we do have

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\Theta}) = 0 \implies \hat{\Theta} \text{ is consistent.}$$

**Example 90.** Let  $\bar{X}$  be the estimator for the population mean  $\mu$ . We can show that  $\bar{X}$  is unbiased, efficient, and consistent!

## 8.2 Maximum Likelihood Estimation

**Definition 91.** Let  $\theta \in \Theta$  be a parameter of a population where  $\Theta$  is the parameter space, and  $\mathbf{x} \in \mathbb{R}^n$  be the realization of the random object  $\mathbf{X} \in \mathbb{R}^n$ ,  $n \in \mathbb{Z}^+$ . The **likelihood function** of  $\mathbf{x}$  is

$$L(\theta) = L(\theta|\mathbf{x}) = \begin{cases} P_{\mathbf{X}|\theta}(\mathbf{X} = \mathbf{x}), & \text{discrete data,} \\ f_{\mathbf{X}|\theta}(\mathbf{x}), & \text{absolutely continuous data.} \end{cases}$$

Then the **maximum likelihood estimator (MLE)** of  $\theta$  is an estimator  $\hat{\theta}$  s.t.

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

$\hat{\theta}$  could be obtained by

1. find out  $L(\theta)$ .
2. take  $\log(L(\theta))$  to obtain the *log-likelihood function*  $l(\theta)$ .
3. partial differentiate  $l(\theta)$  w.r.t.  $\theta$  and equate  $l(\theta)$  to 0.
4. solve for  $\theta$  to obtain the expression for  $\hat{\theta}$ .
5. check that the obtained  $\hat{\theta}$  corresponds to a maximum of the likelihood function by checking if

$$\frac{\partial^2}{\partial \theta^2} l(\hat{\theta}) < 0.$$

6. Change from  $\hat{\theta} = \hat{\theta}(x)$  to  $\hat{\theta} = \hat{\theta}(X)$  so that the final expressions is a *estimator* instead an *estimate*.

### Comments

- The MLE is not necessarily unbiased, i.e. it is possible that

$$\hat{\theta} \neq \theta,$$

where  $\hat{\theta}$  is the MLE and  $\theta$  is the true parameter.

- + The MLE is consistent.

- + The MLE is asymptotically normal, i.e. assume  $(\hat{\theta}_n)_{n \in \mathbb{N}}$  is a consistent sequence of MLEs, then

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow_d N(0, \sigma^2(\theta))$$

for some  $\sigma^2(\theta)$ .

- + The MLE is always asymptotically efficient, and if an efficient estimator exists, it is the MLE.

### 8.3 Confidence Intervals

**Definition 92.** In general, a  $1 - \alpha$  **confidence interval**  $I$  is a random interval that contains the “true” parameter with probability  $\geq 1 - \alpha$ , i.e.

$$P_\theta(\theta \in I) \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

If sample size is large, we can exploit the CLT. Thus, for any desired coverage probability level  $1 - \alpha$  we can define the  $1 - \alpha$  C.I. for  $\theta$  by

$$\left[ \hat{\theta} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \hat{\theta} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right],$$

where  $z_\alpha$  is the  $\alpha$ -quantile of the standard normal.

**Example 93.** A corporation conducts a survey to investigate the proportion of employees who thought the board was doing a good job. 1000 employees, randomly selected, were asked, and 732 said they did. Find a 99% confidence interval for the value of the proportion in the population who thought the board was doing a good job.

We can model each observation as  $X_i \sim \text{Bernoulli}(p)$  for some unknown  $p$ , and we want to find a C.I. for  $p$ , which is also the mean of  $X$ . We have our estimate  $\hat{p} = \bar{x} = 0.732$  for which we use the CLT. Since the variance of  $\text{Bernoulli}(p)$  is  $p(1 - p)$ , we can use  $\bar{x}(1 - \bar{x}) = 0.196$  as an approximate variance. So an approximate 99% C.I. is

$$\left[ 0.732 - 2.576 \times \sqrt{\frac{0.196}{1000}}, 0.732 + 2.576 \times \sqrt{\frac{0.196}{1000}} \right].$$

**Theorem 94.** If  $X_1, X_2, \dots, X_n$  are an i.i.d. sample from  $N(\mu, \sigma^2)$  with  $\sigma$  known, then we can construct an *exact* confidence interval for  $\mu$  as

$$\left[ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right].$$

If  $\sigma$  is unknown, then we can construct an C.I. for  $\mu$  as

$$\left[ \bar{x} - t_{n-1, 1-\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}} \right],$$

where  $t_{\nu, \alpha}$  is the  $\alpha$ -quantile of  $\text{Student}(\nu)$ , which is the Student’s  $t$ -distribution with  $\nu$  degrees of freedom, and  $S_{n-1} = \sqrt{S_{n-1}^2}$  is the bias-corrected sample standard deviation.

#### Comments

- $\text{Student}(\nu)$  is heavier tailed than  $N(0, 1)$  for any number of degrees of freedom  $\nu$ , so the  $t$ -distribution C.I. will always be wider than the Normal distribution C.I. Therefore if we know  $\sigma^2$ , we should use it.

- $\lim_{\nu \rightarrow \infty} \text{Student}(\nu) = N(0, 1)$ .
- For  $\nu > 40$ , the difference between  $\text{Student}(\nu)$  and  $N(0, 1)$  is so insignificant that the  $t$ -distribution is not tabulated beyond this many degrees of freedom, and so there we can instead revert to  $N(0, 1)$  tables.



# Chapter 9

## Hypothesis Testing

### 9.1 Definitions

**Definition 95.** We partition the parameter space  $\Theta$  into two disjoint sets  $\Theta_0$  and  $\Theta_1$  and we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{v.s.} \quad H_1 : \theta \in \Theta_1.$$

We call the  $H_0$  the ***null hypothesis*** and  $H_1$  the ***alternative hypothesis***.

To test the validity of  $H_0$ , we first choose a test statistic  $T(\mathbf{X})$  of the data for which we can find the distribution  $P_T$  under  $H_0$ . Then we identify a rejection region  $R \subset \mathbb{R}$  of low probability values of  $T$  under the assumption that  $H_0$  is true, i.e. a region  $R$  s.t.

$$P(T \in R | H_0) = \alpha$$

for some ***significance level***  $\alpha \in (0, 1)$ . We finally calculate the observed test statistic  $t(\mathbf{x})$  and

- if  $t \in R$  we “reject the null hypothesis at the  $\alpha$  level”.
- if  $t \notin R$  we “do not reject (retain) the null hypothesis at the  $\alpha$  level”.

**Definition 96.** We define the ***p-value*** as

$$p = \sup_{\theta \in \Theta_0} P_{\theta}(\text{observing sth “at least as extreme” as the observation}).$$

**Definition 97.** The **power function** is defined as the mapping

$$\beta : \Theta \mapsto [0, 1].$$

The **power** of a hypothesis test is then defined as

$$\beta(\theta) = P_\theta(\text{reject } H_0).$$

If  $\theta \in \Theta_0$  then we want  $\beta(\theta)$  to be small; If  $\theta \in \Theta_1$  then we want  $\beta(\theta)$  to be large.

The following Figure 9.1 defines the **Type I error** and **Type II error**.

	$H_0$ true	$H_0$ false
do not reject $H_0$	✓	Type II error <i>False negative</i>
reject $H_0$	Type I error <i>False positive</i>	✓ <i>power-function [0,1]</i>

Figure 9.1: Type I and Type II errors

**Example 98.**  $X \sim N(\theta, 1)$ ,  $\theta \in \mathbb{R}$  unknown. We test

$$H_0 : \theta \leq 0 \quad \text{against} \quad H_1 : \theta > 0.$$

Thus  $\Theta = \mathbb{R}$ ,  $\Theta_0 = (-\infty, 0]$ ,  $\Theta_1 = (0, \infty)$ . Suppose we use the critical (rejection) region

$$R = [c, \infty).$$

We choose a critical value  $c$  s.t. the test is of level  $\alpha$ . For  $\theta \leq 0$ :

$$P_\theta(\text{reject } H_0) = P_\theta(X \geq c) = P_\theta(\underbrace{X - \theta}_{\sim N(0,1)} > c - \theta) = 1 - \Phi(c - \theta) \leq 1 - \Phi(c).$$

Thus we choose  $c$  s.t.  $\Phi(c) = 1 - \alpha$ , then  $\forall \theta \in \Theta_0$ ,  $P_\theta(\text{reject } H_0) \leq \alpha$ .

## 9.2 Testing for a Population Mean

Suppose  $X_1, X_2, \dots, X_n$  are i.i.d.  $N(\mu, \sigma^2)$ . We wish to test if  $\mu = \mu_0$  for some specific value  $\mu_0$ , so we can state our null and alternative hypothesis as

$$H_0 : \mu = \mu_0 \quad \text{v.s.} \quad H_1 : \mu \neq \mu_0$$

### 9.2.1 Normal Distribution with Known Variance

Say  $\sigma^2$  is known and  $\mu$  is unknown. We can set up a test statistic

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \Phi.$$

Thus the rejection region  $R$  can be defined as

$$R = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right) \cup \left(z_{1-\frac{\alpha}{2}}, \infty\right) = \left\{z \mid |z| > z_{1-\frac{\alpha}{2}}\right\}$$

s.t.  $P(Z \in R | H_0) = \alpha$ . As such, we reject  $H_0$  at the  $\alpha$  significance level  $\iff$  our observed test statistic satisfies

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in R \quad \text{or} \quad p\text{-value} = 2(1 - \Phi(|z|)) \leq \alpha.$$

### 9.2.2 Normal Distribution with Unknown Variance

Say  $\sigma^2$  is unknown and  $\mu$  is unknown. We can set up a test statistic

$$T = \frac{\bar{X} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \sim t_{n-1}.$$

The rejection region  $R$  thus changes to

$$R = \left\{t \mid |t| > t_{n-1, 1-\frac{\alpha}{2}}\right\}$$

s.t.  $P(T \in R | H_0) = \alpha$ . As such, we reject  $H_0$  at the  $\alpha$  significance level  $\iff$  our observed test statistic satisfies

$$t = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \in R \quad \text{or} \quad p\text{-value} = 2\left(1 - t_{n-1, 1-\frac{\alpha}{2}}\right) \leq \alpha.$$

## 9.3 Testing for Differences in Population Means

Suppose that

- $\mathbf{X} = (X_1, \dots, X_{n_1})$  are i.i.d.  $N(\mu_X, \sigma_X^2)$  with  $\mu_X$  unknown;
- $\mathbf{Y} = (Y_1, \dots, Y_{n_2})$  are i.i.d.  $N(\mu_Y, \sigma_Y^2)$  with  $\mu_Y$  unknown;
- the two samples  $\mathbf{X}$  and  $\mathbf{Y}$  are independent,

and we want to test

$$H_0 : \mu_X = \mu_Y \quad \text{v.s.} \quad H_1 : \mu_X \neq \mu_Y.$$

### 9.3.1 Normal Distribution with Known Variances

If  $\sigma$  is known, we have

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_1}\right), \quad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right) \implies \bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right),$$

thereby setting up test statistic

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \sim \Phi,$$

following up with the investigation of rejection of null hypothesis.

### 9.3.2 Normal Distribution with Unknown Variances

If  $\sigma$  is unknown, we have

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_{n_1+n_2-2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{X} - \bar{Y})}{S_{n_1+n_2-2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

following up with the investigation of rejection of null hypothesis.

## 9.4 Goodness of Fit

### 9.4.1 Chi-square Test

**Definition 99.** To test for *goodness of fit*, i.e. compare the *observed frequency*  $\mathbf{O} = (O_1, \dots, O_k)$  with the *expected frequency*  $\mathbb{E} = (E_1, \dots, E_k)$ , we set up  $H_0 : \theta = \theta_0$  v.s.  $H_1 : \theta \neq \theta_0$  for the value of the unknown parameter, and use the *chi-square statistic*

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}. \quad (\geq 0)$$

If  $H_0$  were true, then the statistic  $\chi^2$  would approximately follow a *chi-square distribution* with  $\nu = k - m - 1$  degrees of freedom.

#### Comments

- $k$  is the number of values (categories) the simple random variable  $X$  can take.
- $m$  is the number of parameters we needed to estimate from the data ( $\dim(\theta)$ ) in order to calculate the  $p_j$ 's.

- E.g. given a sample without specifying the model, the degree of freedom  $\nu = k - 0 - 1 = k - 1$ ; if say the sample is fitted with a Poisson distribution with rate parameter  $\lambda$ , then  $\nu = k - 1 - 1 = k - 2$ .
- For the approximation to be valid, we should have  $\forall j, E_j \geq 5$ . This may require some merging of categories.
- Larger  $\chi^2$  corresponds to larger deviations from the null hypothesis model; if  $\chi^2 = 0$ , observed counts exactly match those expected under  $H_0$ .
- Since  $\chi^2 \geq 0$ , we always perform a one-sided goodness of fit test using the  $\chi^2$  statistic, looking at the upper tail of the distribution, leading to the rejection region  $R$  at  $\alpha$  level being

$$R = \left\{ x^2 \mid x^2 > \chi_{k-m-1, 1-\alpha}^2 \right\}.$$

### 9.4.2 Independence using Chi-square Statistic

Assume two discrete random variables  $X$  and  $Y$  that can each take finite values which are jointly distributed with unknown probability mass function  $p_{XY}$ . To determine if  $X$  and  $Y$  are independent, we can do the following:

Let the ranges of  $X$  and  $Y$  be  $\{x_1, \dots, x_k\}$  and  $\{y_1, \dots, y_l\}$  respectively. Then we can form the following  $k \times l$  **contingency table**:

Table 9.1: contingency table example

	$y_1$	$y_2$	$\dots$	$y_l$	
$x_1$	$n_{11}$	$n_{12}$		$n_{1l}$	$n_{1\bullet}$
$x_2$	$n_{21}$	$n_{22}$		$n_{2l}$	$n_{2\bullet}$
$\vdots$					
$x_k$	$n_{k1}$	$n_{k2}$		$n_{kl}$	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet l}$	$n$

where  $n_{ij}$  represents the number of times we observe the pair  $(x_i, y_i)$ ,  $n_{i\bullet}$  represents the frequencies of  $x_i$  in the sample, and similarly for  $n_{\bullet j}$ . Under the null hypothesis,

$$H_0 : X \text{ and } Y \text{ are independent,}$$

we can compute the expected values of entries of the contingency table by

$$\hat{n}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

since

$$\hat{p}_{i\bullet} = p_X(x_i) = \frac{n_{i\bullet}}{n}, \quad \hat{p}_{\bullet j} = p_Y(y_j) = \frac{n_{\bullet j}}{n} \implies \hat{p}_{ij} = \hat{p}_{i\bullet} \times \hat{p}_{\bullet j} = \frac{n_{i\bullet} n_{\bullet j}}{n^2}$$

and multiply both sides with  $n$  to obtain the desired quantity  $\hat{n}_{ij}$ . We can then set up the chi-square test statistic

$$x^2 = \sum_{i,j} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

with the degrees of freedom  $\nu = kl - (k - 1) - (l - 1) - 1 = (k - 1)(l - 1)$ . Hence the rejection region for a hypothesis test of independence in a  $k \times l$  contingency table at  $\alpha$  level is given by

$$R = \left\{ x^2 \mid x^2 > \chi_{(k-1)(l-1), 1-\alpha}^2 \right\}.$$